Llama-Mimi: 意味・音響トークンを交互配置した音声言語モデル

杉浦一瑳 ♣,◇ 栗田修平 ‡,◇ 小田悠介 ◇ 東中竜一郎 ♡,◇
♣ 京都大学 ◇NII LLMC ‡ 国立情報学研究所 ♡ 名古屋大学
sugiura.issa.q29@kyoto-u.jp

概要

本研究では，音声言語モデルにおけるシンプルなアーキテクチャの可能性を探索することを目的として，意味トークンと音響トークンを交互配置した系列を単一の Transformer デコーダで同時にモデル化する音声言語モデル Llama-Mimi を提案する．多面的な評価の結果，Llama-Mimi は音響一貫性において最先端の性能を達成する一方，言語的能力の課題が明らかになった．また，量子化段数を増やすことによる音響の質と言語的性能のトレードオフを明らかにした．本研究で使用したモデル，コード，音声サンプルは公開している．[1)]

1 はじめに

音声言語モデルは，音声波形を離散トークン列に変換し，それらを自己回帰的に予測することで音声生成を言語モデリングとして扱う [1, 2, 3]．GSLM によって離散トークンを用いた音声モデリングの有効性が示された後 [1]，GSLM に存在していた言語的能力の問題や単一話者の問題が少しずつ改善されている [2]．

しかし，これらの発展は複数のモデルによるマルチステージ化や複雑なトークン配置などを用いており，モデルの複雑性は増大している [4, 5, 6]．例えば AudioLM は，まず単一デコーダで意味トークンを生成した後，複数段のモデルが残りの音響トークンを生成する [2]．Moshi は RQ-Transformer [7] に基づく二重 Transformer 構造を採用し，backbone Transformer がフレーム間の時間方向の系列を処理し，depth Transformer が backbone Transformer の生成トークンを条件つけしてフレーム内のトークン列を処理する [5]．これらの設計により言語的，音響的に強いモデルが開発される一方で，学習・推論が複雑化する．

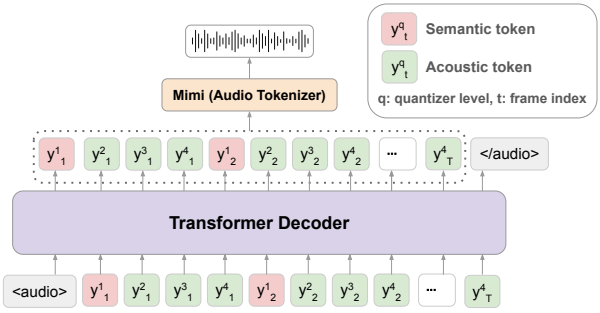

図 1 Llama-Mimi のモデルアーキテクチャ．

これに対し，言語タスクでは単一の Transformer デコーダを用いた大規模言語モデル (LLM) が多様なタスクで成功を収めており [8, 9, 10]，スケーリング則や学習安定性に関する知見も蓄積されている [11, 12]．

本研究では，音声言語モデルにおけるシンプルなアーキテクチャの可能性を探索することを目的として，意味トークンと音響トークンを統合して扱うトークナイザ Mimi [5] とオープンな大規模言語モデル Llama を用いた音声言語モデル Llama-Mimi を提案し，有効性を検証する．Llama-Mimi は，意味トークンと音響トークンを交互配置することで，Transformer decoder が音声トークン系列をテキスト系列と同様に扱う．実験では，Llama-Mimi を尤度ベースと生成ベースで評価を行った．その結果，Llama-Mimi は音響の一貫性を識別するタスクにおいて最先端の性能を示した．一方で，言語的能力の課題があることも明らかになった．アブレーション実験では，量子化段数による言語的一貫性と音響品質のトレードオフを明らかにした．

2 Llama-Mimi

図 1 に Llama-Mimi の概要を示す．Llama-Mimi はニューラル音声コーデックの Mimi [5] を用いて音

1) https://speed1313.github.io/llama-mimi

声波形を残差ベクトル量子化 [13, 14] に変換する. Mimi は，自己教師あり学習により獲得された音声表現モデルである WavLM [15] を用いて，残差ベクトル量子化の第一量子化ベクトルを semantic distillation する統合トークナイザ [16] である.

我々は，単一チャネルの音声波形 $\mathbf{x} \in \mathbb{R}^T$ が与えられたとき，Mimi により音声波形 $\mathbf{x}$ を離散トークン列 $\mathbf{h}$ に変換する:

$$\mathbf{h} = (y_1^1, y_1^2, \ldots, y_1^Q, y_2^1, \ldots, y_{T'}^Q),$$

ここで，Q は量子化段数，T'（$T' \ll T$）は音声のフレーム数，y_t^q は時刻 t における q 階層のトークンを指す. 各フレームは 8 階層のトークンで構成され，第一階層は意味トークン，第二から第八は音響トークンの役割を持つ.

Mimi によって得られた離散音声トークン列を，単一の Transformer デコーダでモデリングする. バックボーンモデルには Meta によって開発された open-weight LLM の Llama 3 [17] を採用する. 各フレームにおいて，モデルはまず意味トークンを予測し，それらに条件付けて音響トークンを続けて生成することで，言語的整合性と細かな音響表現を両立する.

実装においては，Llama 3 が音声トークンを処理できるよう 語彙に音声トークン及び特別トークンとして音声トークン列の先頭，末尾を示す<audio>, </audio>を追加する. 学習時の目的関数には次トークン予測 [8] を用いる. 推論時には従来の LLM と同様にトークンを自己回帰的に生成する. 生成されたトークンは Mimi によって順次デコードされ音声フレームへと変換され，</audio> トークンが出力された時点で終了する. サンプリング時にはトークン制約をかけて Mimi の RVQ の順序（$1 \to 2 \to \cdots \to Q$）に従ったトークン列を生成することが望ましいが，我々の実験ではトークン制約をかけずとも正しい順序で生成が行われた.

3 モデルの学習

モデル 本研究では，Llama 3 [17]1.3B および 8B を用い，それぞれ Llama-Mimi-1.3B，Llama-Mimi-8B と呼ぶ. 系列長と再構成品質のバランスを取るため，1 フレームごとの量子化段数はデフォルトで $Q = 4$ とし，1 秒の音声を 50 トークンで表現する設定とした. また，音声言語モデルのバックボーンのパラメータとして，事前学習済み LLM を用いることの有用性を示した TWIST [3] に従い，モデルは事前学習済み の Llama 3 で初期化した. Mimi のパラメータは固定とした.

データセット モデルの学習データセットには Libri-Light [18]，The People's Speech [19]，VoxPopuli [20]，Emilia [21] を合わせた約 24 万時間からなる大規模英語音声コーパスを用いた. なお，すべてのサンプルは最大 20 秒に切り詰めて学習した. 切り詰め後のデータセットは 1~20 秒の幅広い長さの音声で構成されており，このデータセットを用いることで，モデルは多様な長さの音声を生成できる.

ハイパーパラメータ モデルの学習には AdamW オプティマイザ [22] を用いた. グローバルバッチサイズは 1,024，最大系列長は 1,024 トークンとした. 学習率は Warmup-Stable-Decay（WSD）スケジュール [23] に従い，学習ステップは 100,000 ステップとした. 最大学習率は 3×10^{-4} として 1,500 ステップウォームアップさせた. ウォームアップ後は学習ステップの内 80% のまで最大学習率を維持し，残り 20% で 3×10^{-5} まで線形に減衰させた.

学習は 32 枚の NVIDIA H200 GPU 上で FSDP [24] を用いて実施し，Llama-Mimi-1.3B の学習には約 48 時間を要した.

4 評価

Llama-Mimi の有効性を評価するために，尤度ベースおよび生成ベースの評価を行った.

ベースライン ベースラインとして GSLM [1]，TWIST [3]，Flow-SLM [6]，Moshi [5] を用いた. SSL ベースの GSLM と TWIST-1.3B は，意味情報に関するタスクで高い性能を示す. 一方，条件付きフローマッチング（CFM）に基づく Flow-SLM-1B-ext は音響面で優れた性能を持つ. Moshi は Mimi を音声トークナイザとして用い，時間方向と深さ方向の併用（temporal-plus-depth）に基づく手法を採用している. なお，Moshi の音声のみ事前学習後のチェックポイントは公開されていないため，Moshi [5] で報告されたスコアを引用した.

4.1 尤度ベース評価

ここでは尤度に基づくベンチマークを用いて音響的知識と意味的知識の両面を評価する.

音響タスクには SALMon [25] を用いる. SALMon は音響的一貫性と音響・意味整合性の 2 カテゴリから構成される. 音響的一貫性タスクでは話者や感情

表 1　尤度ベース評価. † は各モデルの原論文で報告されているスコアを示す. **太字**は最良値，下線は次善値を示す. sStoryCloze は sTopic-StoryCloze を示す.

モデル	音声トークナイザ	音響							意味		
		SALMon									
		Consistency					Alignment		sWUGGY	sBLIMP	sStoryCloze
		Sentiment	Speaker	Gender	BG (all)	Room	Sentiment	BG			
GSLM-151M	SSL	64.0	70.0	81.5	68.0	53.5	48.0	55.0	53.9	51.4	67.3
Flow-SLM-1B-ext†	CFM	65.0	76.5	80.0	64.5	73.5	**57.0**	53.0	-	-	-
TWIST-1.3B	SSL	61.5	69.0	69.5	60.5	59.0	53.0	**56.5**	71.7	56.8	69.9
Llama-Mimi-1.3B	Audio Codec	**79.0**	85.0	83.5	**73.5**	**92.0**	48.5	53.5	68.7	54.3	64.0
TWIST-7B	SSL	61.5	71.0	70.0	60.0	61.5	51.0	54.5	73.5	58.7	75.4
Moshi-7B†	Audio Codec	-	-	-	-	-	-	-	**74.3**	**58.9**	**81.8**
Llama-Mimi-8B	Audio Codec	76.5	**86.5**	**85.5**	73.0	**92.0**	46.5	52.5	68.8	55.1	67.6
Human†	-	97.2	91.5	98.6	88.7	94.4	93.3	95.8	-	-	90.2

表 2　プロンプトとそれに対するモデル出力の文字起こしを示す. 文字起こしには Whisper Turbo を用いた.

プロンプト	Llama-Mimi-1.3B	Llama-Mimi-8B
Hi, how are you?	Hi. How are you? I guess this is a thing indifferent to you. When doing the second call, you bring me back to the gate. You know that it wasn't too dark to mine, but when it took you from my side, what? It's been worse.	Hi. How are you? I'm astonished to see you. Why, I've been looking at you and you had put on yourself out pretty well. And what do you suppose I did there last night?
1,2,3,4,5,6	1,2,3,4,5,6,7,9 ominous in the stream thing of manual reading and genius and personal discoveries and daily vehicles and data scientific studies and advice and probably hymns on classical history and biology and all as they entertain their marvelous ideals of vigorous golf competition which	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 9, 8.

などの音響的属性が発話中で不自然に変化したサンプルより，そうでない自然なサンプルに高い尤度を割り当てられるかを評価する. 音響・意味整合性タスクでは意味的内容と背景音や感情などの音響信号が一致した自然なサンプルを，不自然なサンプルより高いスコアを与えられるかを測定する.

意味タスクには sWUGGY [26]，sBLIMP [26]，および sTopic-StoryCloze [3] を使用する. sWUGGY は実在語（例: "oscillation"）と疑似語（例: "odenacia"）のどちらに高い確率を割り当てるかを評価する. sBLIMP は文法的に正しい文（例: "The sweater isn't folding"）を誤った文（例: "The sweaters isn't folding"）より好むかを測定する. sTopic-StoryCloze は音声プロンプトに対して意味的に整合した次の音声を正しく選べるかを評価する.

また，AudioLM [2] に倣い SALMon を除くすべてのタスクで Llama-Mimi の尤度は意味トークンに限定して計算した. なお音響トークンを除外した方が SALMon を除くすべてのベンチマークで性能が向上することを確認した.

結果　表 1 に結果を示す. Llama-Mimi は SALMon の音響的一貫性タスクで最良の性能を達成した. これは過去のすべてのトークンに注意を向け微細な音響変化を捉える表現学習を可能にする本手法の設計によるものと考えられる. 一方で意味タスクでは TWIST モデルに及ばなかった. これはシーケンスが長くなることでグローバルな情報の扱いが難しくなるためと推察される. また，モデルサイズは音響タスクにはほとんど影響しないが意味タスクには明確な改善をもたらすことも確認した.

表 3　話者の一貫性と発話内容の品質. LLM スコアは 1 (低) - 10 (高).

モデル	話者類似度↑	発話内容の質	
		LLM↑	PPL↓
GSLM-151M	0.112	1.98	214.9
TWIST-1.3B	0.110	**3.63**	215.2
Llama-Mimi-1.3B (Ours)	0.346	3.01	165.5
TWIST-7B	0.109	4.38	183.6
Llama-Mimi-8B (Ours)	**0.348**	4.03	**144.0**
参照音声	0.484	6.06	421.2

4.2 生成ベース評価

生成ベース評価では，短い音声プロンプトをモデルに入力して生成される続きの音声について，話者の一貫性および発話内容の品質を評価する．実験には LibriSpeech の test-clean サブセット（2,620 事例）を用いる．各サンプルの先頭 3 秒をプロンプトとして切り出し，最大 20 秒までの生成を行った．サンプリング設定は temperature = 0.8，top-k = 30 とした．

話者一貫性　モデルの生成音声の話者の一貫性を測定するため，pyannote の WavLM ベースの埋め込みモデル [27] を用いて音声埋め込みを抽出し，プロンプトと生成音声のコサイン類似度を計算した．

発話内容の質　GSLM [1] は生成音声の書き起こしに対してパープレキシティを計算することで発話内容の品質を評価する手法を提案した．しかしパープレキシティは系列長やサンプリングに強く依存するため不安定であることが指摘されている [3]．本研究では，発話内容の品質評価のために LLM を用いた柔軟な評価手法の LLM-as-a-Judge [28] を採用する．LLM-as-a-Judge を用いた評価では，音声プロンプトおよび生成音声を Whisper Turbo [29] で文字起こしした後，GPT-4o（`gpt-4o-2024-11-20`）[10] に以下のプロンプトを用いて生成音声を 1 (低) - 10 (高) の尺度で評価させる．

> Given the following prompt and completion, rate the quality of the completion on a scale from 1 to 10, where 10 is the best possible completion. Consider relevance, coherence, fluency, and informativeness. Output only the score as an integer.
> Prompt: {prefix}
> Completion: {suffix}

スコアの決定性を確保するため，temperature は 0 に設定した．実験では生成音声の書き起こしに対して GPT-2 [8] を用いたパープレキシティも併せて報告して LLM-as-a-Judge による評価と比較する．

結果　生成ベースの評価では，TWIST-1.3B および GSLM と比較を行った．Flow-SLM-1B-ext と Moshi については音声のみで事前学習されたチェックポイントが公開されていないため除外した．表 3 に結果を示す．話者一貫性に関しては Llama-Mimi はいずれもベースラインより高いスコアを達成しており，話者性の保持に優れていることが分かる．LLM-as-a-Judge 評価では，Llama-Mimi-1.3B は GSLM-151M と TWIST-1.3B の中間程度のスコアを示し，尤度ベースの意味タスクにおける評価結果と整合性がある．なお，発話内容の質の評価において，パープレキシティスコアは参照音声で最悪のスコアとなっており，問題が示唆される．一方 LLM-as-a-Judge による評価はより一貫しており，モデルサイズが大きいほど高スコアとなり参照音声が最も高いスコアを得ていることから妥当な評価であるといえる．

表 4　量子化段数の効果．音質と LLM スコアは 1 (低) - 10 (高)．

量子化段数	音質				話者類似度	発話内容の質
	PQ	PC	CE	CU		LLM
2	5.02	**1.62**	4.58	4.72	0.201	**3.53**
4	5.55	**1.62**	5.09	5.26	0.346	3.01
8	**6.01**	1.61	**5.40**	**5.66**	**0.474**	2.54
参照音声	5.72	1.45	5.43	5.53	0.484	6.06

4.3 生成事例分析

TWIST のデモ [3] で用いられた音声プロンプトを用いて Llama-Mimi の出力を事例分析した．表 2 に示すように Llama-Mimi-8B は Llama-Mimi-1.3B より自然な生成音声であり，モデルが大きいほど言語能力に優れていることがわかる．

4.4 量子化段数の効果

ここでは，量子化段数が音質，話者類似度，および音声内容品質に与える影響を評価した．音質評価には Audiobox-Aesthetics [30] を用い，制作品質 (PQ)，制作の複雑さ (PC)，内容の楽しさ (CE)，内容の有用性 (CU) の 4 指標でスコア化する．生成ベース評価の実験同様，プロンプト用のデータセットに LibriSpeech test-clean [31] を用い，量子化段数を $Q \in \{2,4,8\}$で変化させる．モデルは Llama 1.3B を使用した．

表 4 に量子化段数の効果分析の結果を示す．量子化段数を増やすと音質と話者類似度は向上する一方，発話内容の質は低下することがわかる．

5 おわりに

本研究では Mimi と Llama を用いたシンプルな音声言語モデル Llama-Mimi を提案した．実験では，Llama-Mimi は音響一貫性タスクで最先端の性能を示した一方，言語的能力の課題を示した．さらに，量子化段数による音響的能力と言語的能力のトレードオフを明らかにした．

謝辞

本研究では，産総研及び AIST Solutions が提供する ABCI 3.0 を「ABCI 3.0 開発加速利用」を支援を受けて利用した.

参考文献

[1] Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, et al. On generative spoken language modeling from raw audio. **ACL**, 2021.

[2] Zalán Borsos, Raphaël Marinier, Damien Vincent, et al. AudioLM: A language modeling approach to audio generation. **IEEE/ACM TASLP**, 2023.

[3] Michael Hassid, Tal Remez, Tu Anh Nguyen, et al. Textually pretrained speech language models. In **NeurIPS**, 2023.

[4] Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, et al. On the landscape of spoken language models: A comprehensive survey. **arXiv preprint arXiv:2504.08528**, 2025.

[5] Alexandre Défossez, Laurent Mazaré, Manu Orsini, et al. Moshi: a speech-text foundation model for real-time dialogue. **arXiv preprint arXiv:2410.00037**, 2024.

[6] Ju-Chieh Chou, Jiawei Zhou, and Karen Livescu. Flow-SLM: Joint learning of linguistic and acoustic information for spoken language modeling. **arXiv preprint arXiv:2508.09350**, 2025.

[7] Doyup Lee, Chiheon Kim, Saehoon Kim, et al. Autoregressive image generation using residual quantization. In **CVPR**, 2022.

[8] Alec Radford, Jeffrey Wu, Rewon Child, et al. Language models are unsupervised multitask learners. **OpenAI blog**, 2019.

[9] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In **NeurIPS**, 2020.

[10] OpenAI. GPT-4o system card. **arXiv preprint arXiv:2410.21276**, 2024.

[11] Jared Kaplan, Sam McCandlish, Tom Henighan, et al. Scaling laws for neural language models. **arXiv preprint arXiv:2001.08361**, 2020.

[12] Mitchell Wortsman, Peter J. Liu, Lechao Xiao, et al. Small-scale proxies for large-scale transformer training instabilities. 2023.

[13] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, et al. SoundStream: An end-to-end neural audio codec. **IEEE/ACM TASLP**, 2021.

[14] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. **TMLR**, 2023.

[15] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. **IEEE**, Vol. 16, No. 6, pp. 1505–1518, 2022.

[16] Xin Zhang, Dong Zhang, Shimin Li, et al. SpeechTokenizer: Unified speech tokenizer for speech language models. In **ICLR**, 2024.

[17] AI@Meta. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[18] Jacob Kahn, Morgane Rivière, Weiyi Zheng, et al. Libri-Light: A benchmark for asr with limited or no supervision. In **ICASSP**, 2020.

[19] Daniel Galvez, Greg Diamos, Juan Ciro, et al. The People's Speech: A large-scale diverse english speech recognition dataset for commercial usage. In **NeurIPS**, 2021.

[20] Changhan Wang, Morgane Riviere, Ann Lee, et al. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In **ACL**, 2021.

[21] Haorui He, Zengqiang Shang, Chaoren Wang, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. **arXiv preprint arXiv:2407.05361**, 2024.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **ICLR**, 2019.

[23] Kaiyue Wen, Zhiyuan Li, Jason S. Wang, et al. Understanding warmup-stable-decay learning rates: A river valley loss landscape view. In **ICLR**, 2025.

[24] Yanli Zhao, Andrew Gu, Rohan Varma, et al. PyTorch FSDP: Experiences on scaling fully sharded data parallel. **arXiv preprint arXiv:2304.11277**, 2023.

[25] Gallil Maimon, Amit Roth, and Yossi Adi. Salmon: A suite for acoustic language model evaluation. **arXiv preprint arXiv:2409.07437**, 2025.

[26] Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, et al. The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. **arXiv preprint arXiv:2011.11588**, 2020.

[27] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, et al. pyannote.audio: neural building blocks for speaker diarization. In **ICASSP**, 2020.

[28] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. **arXiv preprint arXiv:2306.05685**, 2023.

[29] Alec Radford, Jong Wook Kim, Tao Xu, et al. Robust speech recognition via large-scale weak supervision. In **ICML**, 2023.

[30] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, et al. Meta Audiobox Aesthetics: Unified automatic quality assessment for speech, music, and sound. **arXiv preprint arXiv:2502.05139**, 2025.

[31] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In **ICASSP**, 2015.

[32] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. **IEEE/ACM TASLP**, 2021.

[33] Amitay Sicherman and Yossi Adi. Analysing discrete self supervised speech representation for spoken language modeling. In **ICASSP**, 2023.

[34] Eugene Kharitonov, Ann Lee, Adam Polyak, et al. Text-free prosody-aware generative spoken language modeling. In **ACL**, 2022.

[35] Jade Copet, Felix Kreuk, Itai Gat, et al. Simple and controllable music generation. **arXiv preprint arXiv:2306.05284**, 2024.

[36] Dong Zhang, Shimin Li, Xin Zhang, et al. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In **EMNLP**, 2023.

[37] Soumi Maiti, Yifan Peng, Shukjae Choi, et al. VoxtLM: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In **ICASSP**, 2024.

[38] Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, et al. AudioPaLM: A large language model that can speak and listen. **arXiv preprint arXiv:2306.12925**, 2023.

[39] Tu Anh Nguyen, Benjamin Muller, Bokai Yu, et al. SpiRit-LM: Interleaved spoken and written language model. **ACL**, 2025.

[40] Yongxin Zhu, Dan Su, Liqiang He, et al. Generative pre-trained speech language model with efficient hierarchical transformer. In **ACL**, 2024.

[41] Dongchao Yang, Jinchuan Tian, Xu Tan, et al. UniAudio: Towards universal audio generation with large language models. In **ICML**, 2024.

A　関連研究

音声言語モデルは急速に発展している [1, 2, 3]. GSLM [1] は自己教師あり音声表現 [32] を k-means により離散化し，Transformer デコーダで次トークン予測を行うことで音声生成を実現した先駆的研究である．しかし音響品質には限界があり，これらの表現は音素情報と高い相関を持つ一方話者性や性別とは弱い相関しか持たない [33]. pGSLM [34] は音素および韻律情報を取り入れたマルチストリーム Transformer によりこの枠組みを拡張した．

AudioLM [2] はニューラル音声コーデック [13, 14] によって得られる意味・音響トークンを階層的に生成することで音質と長期一貫性を大幅に改善した．TWIST [3] は事前学習済みテキスト LM による初期化が性能向上と収束高速化に寄与することを示した．MusicGen [35] はマルチストリームトークンの遅延生成による性能変化を分析した．近年では離散音声トークンを LLM に組み込むことで単一の Transformer デコーダで ASR や TTS を含む多様な音声・テキストタスクを扱うモデルも登場している [36, 37, 38]．Spirit-LM [39] は単語レベルのテキストと音声トークンを交互配置し，モダリティ間の同時生成を実現した．また Moshi [5] は RQ-Transformer [7] により，マルチストリームトークンを用いた双方向対話をモデル化している．

複数ストリームからなる意味トークンと音響トークンを扱うため各種生成戦略が提案されている [4]. Coarse-to-fine アプローチでは，まず全意味トークン列を生成した後音響トークンを条件付きで生成する [2]. 高い性能を示す一方でストリーミングには適さない．

Temporal-plus-depth アプローチ [40, 41, 5] はバックボーン Transformer がフレーム間依存を Depth Transformer がフレーム内の複数トークンを予測する入れ子構造を採用する．ただしバックボーン Transformer はフレームレベルで集約された情報のみを入力とするため，各フレーム内の個々のトークンに直接注意できず性能上の制約となりうる．

Interleaved coarse-and-fine アプローチ [39] では意味・音響トークンを時間軸で整列させ，各フレームで交互配置することで，単一モデルによる同時処理を可能とする．Spirit-LM [39] はこの戦略の例であるが，意味・ピッチ・スタイルといった複数のトークナイザを用いている．